DOCUMENT RESUME

ED 422 355 TM 028 907

AUTHOR Plake, Barbara S.; Impara, James C.; Spies, Robert; Hertzog, Melody; Giraud, Gerald

TITLE Setting Performance Standards on Polytomously Scored Assessments: An Adjustment to the Extended Angoff Method.

PUB DATE 1998-04-00

NOTE 11p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Diego, CA, April 13-17, 1998).

PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS *Constructed Response; *Cutting Scores; High School Students; High Schools; Minimum Competencies; *Prediction; *Scoring; *Standards; Writing Tests

IDENTIFIERS *Angoff Methods; *Polytomous Scoring; Standard Setting

ABSTRACT

Setting performance standards on constructed-response assessments involving polytomously scored exercises presents a challenge for measurement practitioners. Some standard setting methods designed for use with multiple-choice, dichotomously scored assessments entail aggregating item performance estimates across a panel of experts. For these items, the experts are asked to predict the probability that a minimally competent candidate will correctly answer each of the items in the test. When working with constructed-response, polytomously scored assessments, panelists are often asked to predict the score that would be obtained by a minimally competent candidate and these expected score values are aggregated to determine the passing score. The resultant cutscore often has been found in practice to be unrealistically high. This study investigates the effectiveness of an adjustment technique to reduce the possible inflation of cutscores. Candidates are asked to estimate the proportion of minimally competent candidates who will answer the item (or pass the examination) correctly, and proportions are used as weights in computing the adjusted minimum passing score. The study applied the adjusted extended Angoff approach to a high school writing assessment involving 23 teachers. Application of the adjustment procedure was less than successful for a variety of reasons. The adjustment was minimal and panelists felt that it was unnecessary. In addition, the ramifications of revealing Round 2 results in order to gather these adjustments had negative consequences. Research is needed to study other possible adjustment strategies. (Contains one table and six references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 422 355

Setting Performance Standards on Polytomously Scored Assessments: An Adjustment to the Extended Angoff Method

Barbara S. Plake, James C. Impara, Robert Spies,
Melody Hertzog, Gerald Giraud

University of Nebraska-Lincoln

Running Head: Adjustment to Extended Angoff

Paper presented at the 1998 annual meeting of the American Educational Research Association Annual Meeting, San Diego, CA

TM028907

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Barbara Plake
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Setting Performance Standards on Polytomously Scored Assessments: An Adjustment to the Extended Angoff Method

Abstract

Setting performance standards on constructed-response assessments involving polytomously scored exercises presents a challenge for measurement practitioners. Some standard setting methods designed for use with multiple-choice, dichotomously scored assessments entail aggregating item performance estimates across a panel of experts. For these items, the experts are asked to predict the probability that a minimally competent candidate will correctly answer each of the items in the test. When working with constructed-response, polytomously scored assessments, panelists are often asked to predict the score that would be obtained by a minimally competent candidate and these expected score values are aggregated to determine the passing score. The resultant cutscore often has been found in practice often to be unrealistically high. This study investigates the effectiveness of an adjustment technique to reduce the possible inflation of cutscores.

Setting Performance Standards on Polytomously Scored Assessments: An Adjustment to the Extended Angoff Method

Introduction

Setting passing scores entails determining the minimum passing score on an assessment. Most standard setting methods were developed specifically for multiple-choice tests. Judgmental standard setting methods, like the Angoff (1971) method, are the most prevalent methods used in licensure and certification fields (Sireci & Biskin, 1992). With the Angoff method, panelists are asked to make item performance predictions for a randomly selected, minimally competent candidate (MCC). With dichotomously scored items, this is often operationalized as predicting the proportion of MCCs who would be able to answer the item right (or get a score of 1). Item performance estimates are aggregated across items, yielding an implicit compensatory cutscore for each of the panelists. Panelists' cutscores are then averaged to determine the minimum passing score for the test.

Setting passing scores with polytomously scored, constructed response assessments is a serious challenge for educational measurement practitioners. The methods developed for multiple-choice tests, consisting of a large number of items each scored dichotomously, do not generalize easily to situations involving polytomously scored, constructed response tests.

The most prevalent practice in the field of licensure and certification for setting passing scores on polytomously scored, constructed response tests is a variation of an Angoff methodology (Plake, 1996). Under this paradigm, panelists are asked to estimate, for each exercise comprising the assessment, the score that would be obtained by a randomly selected MCC. These minimum

passing scores per exercise are then aggregated to yield the panelist's passing score . The panelists' minimum passing scores are then averaged to yield the overall minimum passing score for the assessment. This approach has been called the "extended Angoff method" (Hambleton & Plake, 1996).

Aggregating the individual exercises' cutscores, which is the basis for setting the overall passing score on many constructed-response assessments sometimes results in the final passing score that is unrealistically high. Certification agencies using this method report that the impact of applying these cutscores results in too few candidates passing; further, validity studies often verify that qualified candidates who should has passed the test have scores lower than the cutscore when the extended Angoff approach is used (Plake, 1996).

One reason for this effect, sometimes called the "Cascading Effect" (Plake, 1996), is due to the less that perfect correlations between candidate performance on the questions that comprise the test. Linn and Shepard (1997) have shown that when panelists routinely set the performance standard for individual questions above the mean of the question's score distribution, the aggregate effect is that fewer examinees pass the examination than would have been the case with perfectly correlated questions. The degree of impact of this effect is a function of the number of questions on the test and the degree of correlation of performance across the questions. The larger the number of questions and the lower the correlation, the greater the impact.

The purpose of this paper was to investigate to utility of a strategy designed to reduce the "Casdading Effect" on the proportion of candidates passing a certification examination. When used with dichotomously scored items, panelists are asked to estimate the proportion of minimally competent candidates who will answer the item correctly (or pass the item). Similarly, using the established MPS for the question derived from the standard setting process as

the "passing score for the question", panelists are asked to estimate the proportion of minimally competent candidates who would "pass" the question. These proportions would then be used as weights in computing the adjusted minimum passing score (Norcini, Stillman, Sutnick, Regan, Haley, Williams, & Friedman, 1993).

Method

This study applied this adjusted extended Angoff approach to a high school level writing assessment. This assessment is part of a larger criterion-referenced assessment program at a large metropolitan school district in the midwest. The purpose of the assessment program is to identify students who could benefit from additional educational support. The assessment program spans grades and content areas; only the high school level writing assessment was chosen for this project.

Instrument. The assessment consists of one writing prompt which is scored on six traits using trait-specific five-point rubrics. Scorers are trained to apply the rubric to the student essays which are written to the prompt "Describe an important person in your life". The six traits are conventions, voice, word choice, organization, sentence fluency, and ideas and content.

Procedure. This study was undertaken during an operational standard setting workshop involving 23 teachers all of whom taught at least one section of tenth grade English. All of the high schools in the district were represented on the panel. During the operational phase of the workshop, panelists were informed of the purpose of the standard setting process, were given an orientation to the process, participated in a discussion of the traits, and identified the knowledge, skills, and abilities of a "Just Competent Student" (JCS) in high school writing for each of the individual traits.

As a means of identifying the expected performance on the traits by the JCS, panelists undertook a paper selection strategy. For each trait, a set of 10 student papers (called "Benchmark Papers"), were identified, 2 illustrative papers for each of the 5 score points. The panelists focused on one trait at a time; traits were assigned to panelists in such a way that each panelists evaluated only three themes. Panelists were directed to select from the set of benchmark papers the two that either represented or bracketed the work of a just competent student. Panelists were not aware of the actual scores for the papers.

Panelists participated in 2 rounds of paper selection. After Round 1, the panelists' initial paper selection choices were analyzed and minimum passing scores for each trait and for the total across all traits were determined. Panelists were informed of these initial minimum passing scores and information about actual student performance on the traits and total score, including the percentage of students in the district who would qualify for additional educational programming if the Round 1 cutscore for the total score was adopted. Following discussion, panelists were given the opportunity to select different student papers (Round 2), if they felt this was appropriate, for each of their assigned traits. The panelists' Round 2 paper choices were used to determine the Round 2 cutscores for each of the six traits and for the total. An evaluation of the standard setting workshop, through Rounds 1 and 2, was then administered.

At this point, panelists were given the same type of impact data for their Round 2 as was provided after Round 1. They were then asked to estimate the proportion of JCSs who would have scores at or above the individual trait cutpoints derived from their Round 2 results. Panelists were informed that these proportions could be used to make adjustments in the final minimum passing value for the writing assessment. An evaluation was administered to gather the panelists' perceptions of the utility of this adjustment technique.

Results

Table 1 shows the results from Rounds 1and 2 for each of the 6 traits. Also shown in Table 1 is the average of the panelists' estimates, for each trait, of the percent of JCSs who would score at or above the minimum passing score set by the Round 2 results. The Total Cut Score, using the Round 1 results, would be set at 13.00. Using the Round 2 results, the Total Cut Score was calculated to be 14.15. When the panelists were asked to estimate the percent of JCS's whose scores would be at or above each of the Round 2 passing scores set for the 6 writing traits, these values ranged from a high of 94.39% to a low of 90.39%, with an average of 93.57%. Therefore, only minimal adjustments were made by the panelists when they estimated the proportion of JCSs who would score at or above the individual trait and total cutpoints. The adjusted overall cutscore was 13.24.

Evaluations indicated that some of the panelists found the process of estimating these proportions confusing and counterproductive to the process, as they felt that they had sufficiently focused on the expected performance of the JCS during the paper selection process in Rounds 1 and 2 of the standard setting process. In addition, an unanticipated outcome occurred. In order to gather the panelists' perceptions of the proportion of JCSs who would score at or above the trait and total cutpoints, the results from Round 2 was revealed to them. In a traditional standard setting study, the Round 2 results (with a range of appropriate values) would be those recommended to the Board for their consideration in setting the final cutscores for the assessment. Most often, the final results are not revealed to the panelists for a variety of reasons, including a desire to keep the final results secure because the Board often decides to alter these cutpoints for psychometric or political reasons. It is considered

compromising if the results of the standard setting study are made public prior to Board consideration. However, because the final adjustment stage was dependent on the Round 2 results, these values were shared with the panelists. The panelists did not maintain silence when the workshop concluded and the results were public knowledge before the Board of Education had an opportunity to consider the policy decision.

Conclusions

Application of an adjustment procedure to an extended Angoff methodology in a school setting was less than successful for a variety of reasons. The adjustment was minimal and the panelists felt it was unnecessary. In addition, the ramifications of revealing the Round 2 results in order to gather these adjustments had negative consequences. It is not recommended that this approach be applied in standard setting situation involving teachers or where public knowledge of the results prior to Board consider could compromise the Board's deliberations. Further research is needed to study other possible adjustment strategies, including obtaining a priori expectations by panelists of the distribution of candidate scores across the score points for the exercises.

References

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R.L. Thorndike (Ed.), Educational Measurement, (2nd ed.). Washington, DC: American Council on Education.

Hambleton, R.K., & Plake, B.S. (1995). Using an extended Angoff procedure to set standards on complex performance assessments. Applied Measurement in Education, 8(1), 41-55.

Linn, R.L. & Shepard, L.A. (June, 1997). Item-by-item standard setting: Misinterpretations of judge's intentions due to less than perfect item intercorrelations. Paper presented at the Large Scale Assessment Conference, Council of Chief State School Officers, Colorado Springs, CO.

Norcini, J.J., Stillman, P.L., Sutnick, A.I., Regan, M.B., Haley, H.L., Williams, R.G., & Friedman, M. (1993). Scoring and standard setting with standardized patients, Evaluation & the Health Professions, 16(3), 322-332.

Plake, B.S. (November, 1996). Setting performance standards for licensure and certification tests, Paper presented at the invitational conference of the National Academy of Sciences on Standard Setting an NAEP, Washington, DC.

Sireci, S.A., & Biskin, B.H. (1992). Measurement practices in national licensing examination programs: A survey. CLEAR Exam Review, III(I), 21-25.

Table 1. Results form Rounds 1 and 2 for each trait, panelists's estimated percentage of Just Competent Students who will attain a passing score or higher on each trait, and adjusted cutscores.

	Round		Estimate	Adjusted
Trait	1	2	% Attaining	Cut
Organization	2.59	2.82	94.39	2.66
Conventions	2.04	2.25	90.74	2.04
Ideas & Conv	1.82	2.18	90.57	1.97
Word Choice	2.50	2.77	90.39	2.50
Voice	1.92	1.92	95.87	1.84
Sent. Fluency	2.13	2.21	93.35	2.06
Total Cut Score	13.00	14.15		13.24